01TXFSM - Machine Learning and Deep Learning

Final Project
First Person Action Recognition

Eros Fanì - s269781
Politecnico di Torino
eros.fani@studenti.polito.it

Gabriele Trivigno - s276807
Politecnico di Torino
gabriele.trivigno@studenti.polito.it

Cristiano Gerbino - s277058
Politecnico di Torino
s277058@studenti.polito.it

Abstract

1. Introduction

1.1. Goals

The first goal of the project is to replicate some of the experiments performed in [2] and [1]. The objective of these studies is the First Person Action Recognition: they tried to implement a deep learning model capable to extract meaningful features to automatically predict the action filmed by a wearable camera.

After having replicated these experiments we performed a grid search on the experiments to find the best set of values for the hyperparameters.

At last we have tried to improve the performances of the results of [2] and [1] with some innovative ideas.

1.2. Data exploration

The dataset under analysis is a modified version of GTEA61[1]. The dataset contains the videos in form of frames, and also two kind of preprocessed images: *motion maps* and *optical flows*. The folder schema of the dataset is shown in Figure 1. Videos represent 61 class actions performed by 4 different users (*S1*, *S2*, *S3*, *S4*). Sometimes for some actions more than one video is available. The total number of videos in the dataset is, however, 457, which actually means that it is a quite small dataset.

The optical flow methods try to calculate the motion between two image frames which are taken at times t and $t + \Delta t$ at every voxel position. The warp flow methods try also to remove the motion of the wearable camera. We have two kind of these last representations in our dataset: one computed in the horizontal axis (folder *flow_x_processed*) and one other computed in the vertical axis (folder *flow_y_processed*).

The motion maps are special black-and-white images which represent the spatial location in which the Motion Segmentation task of [1] focuses its attention per each frame. The mmaps present large similarities with the warp flows.

[1] Georgia Tech Egocentric Activity Datasets: http://cbs.ic.gatech.edu/fpv/

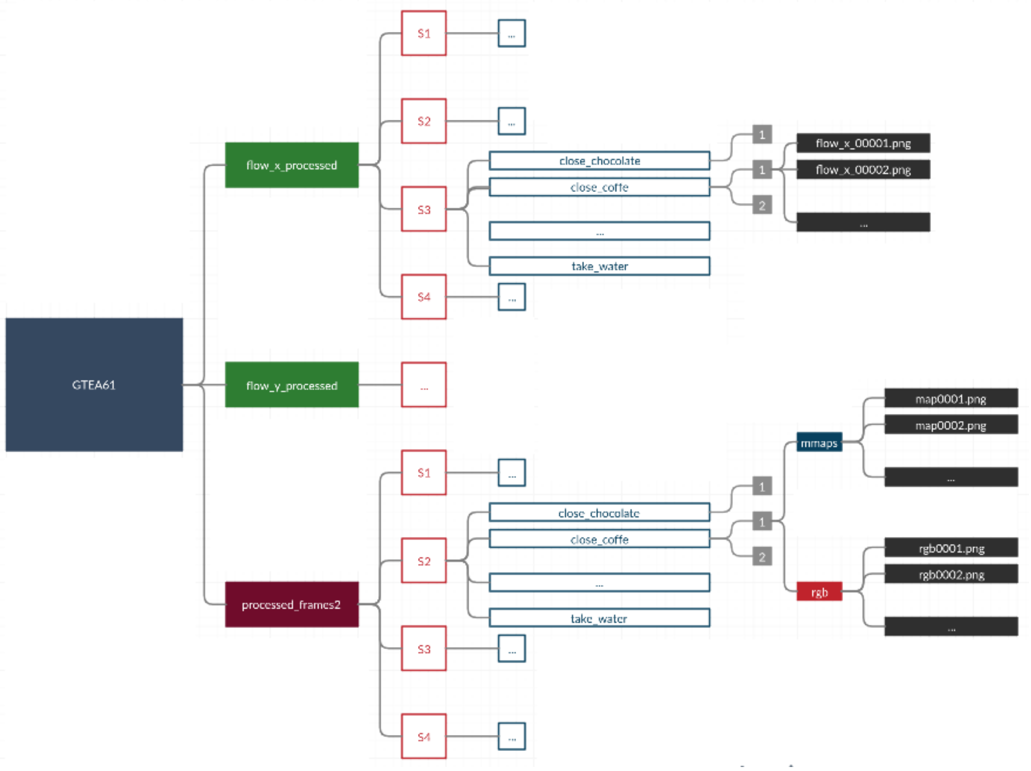

Figure 1: Folder schema of our GTEA61

The differences between the kind of available images in our dataset are shown in Figure 2.

1.3. Data cleaning

References

[1] M. Planamente, A. Bottino, and B. Caputo. Joint encoding of appearance and motion features with self-supervision for first person action recognition, 2020.

[2] S. Sudhakaran and O. Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition, 2018.

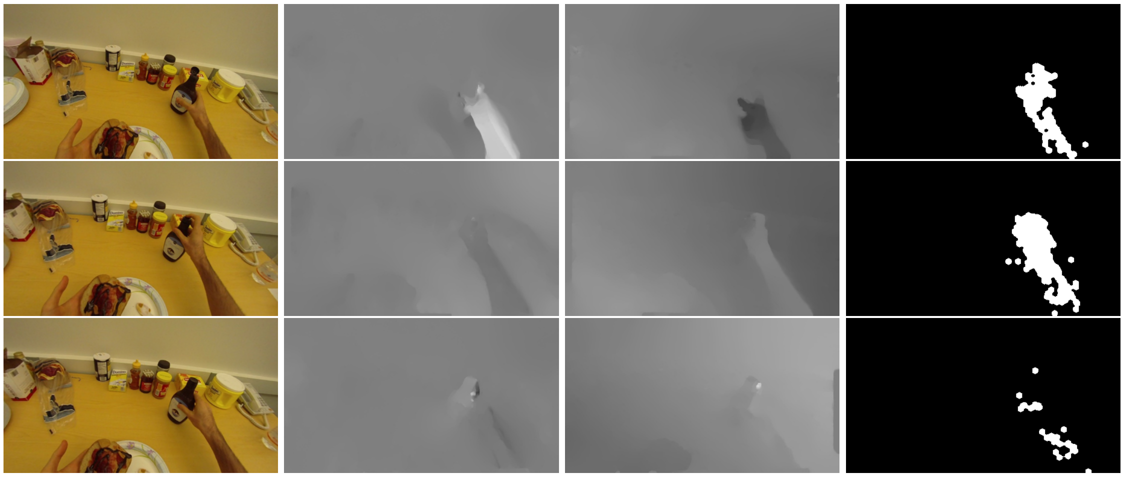

Figure 2: Types of images in our dataset. In this example is shown a sample of images from the *close_chocolate* action. From the left column to the right column: rgbs, warp flows x, warp flows y, motion maps